Improved Signal-to-Noise Ratio Estimation for Speech Enhancement

Cyril Plapous, *Member, IEEE*, Claude Marro, and Pascal Scalart, *Member, IEEE*

***Abstract*—This paper addresses the problem of single-microphone speech enhancement in noisy environments. State-of-the-art short-time noise reduction techniques are most often expressed as a spectral gain depending on the signal-to-noise ratio (SNR). The well-known decision-directed (DD) approach drastically limits the level of musical noise, but the estimated *a priori* SNR is biased since it depends on the speech spectrum estimation in the previous frame. Therefore, the gain function matches the previous frame rather than the current one which degrades the noise reduction performance. The consequence of this bias is an annoying reverberation effect. We propose a method called two-step noise reduction (TSNR) technique which solves this problem while maintaining the benefits of the decision-directed approach. The estimation of the *a priori* SNR is refined by a second step to remove the bias of the DD approach, thus removing the reverberation effect. However, classic short-time noise reduction techniques, including TSNR, introduce harmonic distortion in enhanced speech because of the unreliability of estimators for small signal-to-noise ratios. This is mainly due to the difficult task of noise power spectrum density (PSD) estimation in single-microphone schemes. To overcome this problem, we propose a method called harmonic regeneration noise reduction (HRNR). A nonlinearity is used to regenerate the degraded harmonics of the distorted signal in an efficient way. The resulting artificial signal is produced in order to refine the *a priori* SNR used to compute a spectral gain able to preserve the speech harmonics. These methods are analyzed and objective and formal subjective test results between HRNR and TSNR techniques are provided. A significant improvement is brought by HRNR compared to TSNR thanks to the preservation of harmonics.**

***Index Terms*—*A posteriori* signal-to-noise ratio (SNR), *a priori* SNR, harmonic regeneration, noise reduction, speech enhancement.**

I. Introduction

The problem of enhancing speech degraded by additive noise, when only a single observation is available, has been widely studied in the past and is still an active field of research. Noise reduction is useful in many applications such as voice communication and automatic speech recognition where efficient noise reduction techniques are required.

Scalart and Vieira Filho presented in [1] a unified view of the main single microphone noise reduction techniques where the noise reduction process relies on the estimation of a short-time spectral gain, which is a function of the *a priori* signal-to-noise ratio (SNR) and/or the *a posteriori* SNR. They also emphasize the interest of estimating the *a priori* SNR thanks to the decision-directed (DD) approach proposed by Ephraïm and Malah in [2]. Cappé analyzed the behavior of this estimator in [3] and demonstrated that the *a priori* SNR follows the shape of the *a posteriori* SNR with a frame delay. Consequently, since the spectral gain depends on the *a priori* SNR, it does not match the current frame, and thus the performance of the noise suppression system is degraded.

We propose a method, called two-step noise reduction (TSNR), to refine the estimation of the *a priori* SNR which removes the drawbacks of the DD approach while maintaining its advantage, i.e., highly reduced musical noise level. The major advantage of this approach is the suppression of the frame delay bias leading to the cancellation of the annoying reverberation effect characteristic of the DD approach.

Furthermore, one major limitation that exists in classic short-time suppression techniques, including the TSNR, is that some harmonics are considered as noise only components and consequently are suppressed by the noise reduction process. This is inherent to the errors introduced by the noise spectrum estimation which is a very difficult task for single channel noise reduction techniques. Note that in most spoken languages, voiced sounds represent a large amount (around 80%) of the pronounced sounds. Then it is very interesting to overcome this limitation. For that purpose, we propose a method, called harmonic regeneration noise reduction (HRNR), that takes into account the harmonic characteristic of speech. In this approach, the output signal of any classic noise reduction technique (with missing or degraded harmonics) is further processed to create an artificial signal where the missing harmonics have been automatically regenerated. This artificial signal helps to refine the *a priori* SNR used to compute a spectral gain able to preserve the harmonics of the speech signal.

These two methods, TSNR and HRNR, have been presented in [4] and [5], respectively. This paper is an extension of this previous work. These two approaches are fully analyzed and comparative results are given. They consist in objective evaluation using the cepstral distance and the segmental SNR and subjective evaluation.

This paper is organized as follows. In Section II, we present the parameters and rules of speech enhancement techniques. In Section III, we introduce a tool useful to analyze the SNR estimators. In Section IV, we recall the principle of the DD approach and analyze it. In Section V, we present and analyze the TSNR approach. In Section VI, we describe and analyze the HRNR technique. Finally, in Section VII, we demonstrate the improved performance of the HRNR, compared to TSNR.

Manuscript received April 29, 2005; revised October 4, 2005. This work was supported by France Télécom. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

C. Plapous and C. Marro are with France Télécom R&D–TECH/SSTP, 22307 Lannion Cedex, France (e-mail: cyril.plapous@francetelecom.com; claude.marro@francetelecom.com).

P. Scalart is with University of Rennes—IRISA/ENSSAT, 22305 Lannion, France (e-mail: pascal.scalart@enssat.fr).

Digital Object Identifier 10.1109/TASL.2006.872621

II. Noise Reduction Parameters and Rules

In the usual additive noise model, the noisy speech is given by $x(t) = s(t) + n(t)$, where $s(t)$ and $n(t)$ denote the speech and noise signal, respectively. Let $S(p,k)$, $N(p,k)$, and $X(p,k)$ represent the kth spectral component of the short-time frame p of the speech signal $s(t)$, noise $n(t)$, and noisy speech $x(t)$, respectively. The objective is to find an estimator $\hat{S}(p,k)$ which minimizes the expected value of a given distortion measure conditionally to a set of spectral noisy features. Since the statistical model is generally nonlinear, and because no direct solution for the spectral estimation exists, we first derive an SNR estimate from the noisy features. An estimate of $S(p,k)$ is subsequently obtained by applying a spectral gain $G(p,k)$ to each short-time spectral component $X(p,k)$. The choice of the distortion measure determines the gain behavior, i.e., the tradeoff between noise reduction and speech distortion. However, the key parameter is the estimated SNR because it determines the efficiency of the speech enhancement for a given noise power spectrum density (PSD).

Most of the classic speech enhancement techniques require the evaluation of two parameters: the *a posteriori* SNR and the *a priori* SNR, respectively defined by

$$\mathrm{SNR}_{\mathrm{post}}(p,k) = \frac{|X(p,k)|^2}{\mathrm{E}\left[|N(p,k)|^2\right]} \tag{1}$$

and

$$\mathrm{SNR}_{\mathrm{prio}}(p,k) = \frac{\mathrm{E}\left[|S(p,k)|^2\right]}{\mathrm{E}\left[|N(p,k)|^2\right]} \tag{2}$$

where E[.] is the expectation operator. We define another parameter, the *instantaneous* SNR, as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{inst}}(p,k) &= \frac{|X(p,k)|^2 - \mathrm{E}\left[|N(p,k)|^2\right]}{\mathrm{E}\left[|N(p,k)|^2\right]}\\ &= \mathrm{SNR}_{\mathrm{post}}(p,k) - 1\end{aligned} \tag{3}$$

which can be interpreted as a direct estimation of the local *a priori* SNR in a spectral subtraction approach [6]. Actually, this parameter is useful to evaluate the accuracy of the *a priori* SNR estimator. In practical implementations of speech enhancement systems, the PSDs of speech $\mathrm{E}[|S(p,k)|^2]$ and noise $\mathrm{E}[|N(p,k)|^2]$ are unknown since only the noisy speech spectrum $X(p,k)$ is available. Thus, both the *a posteriori* SNR and the *a priori* SNR have to be estimated. The estimation of the noise PSD $\mathrm{E}[|N(p,k)|^2]$, noted $\hat{\gamma}_n(p,k)$, will not be described in the paper. It can be practically estimated during speech pauses using a classic recursive relation [1] or continuously using the minimum statistics [7] or the minima controlled recursive averaging approach [8] to get a more accurate estimate in case of noise level fluctuations.

Then, the spectral gain $G(p,k)$ is obtained by the function

$$G(p,k) = g\left(\hat{\mathrm{SNR}}_{\mathrm{prio}}(p,k), \hat{\mathrm{SNR}}_{\mathrm{post}}(p,k)\right) \tag{4}$$

depending on the chosen distortion measure. The function g can be chosen among the different gain functions proposed in the literature (e.g., amplitude or power spectral subtraction, Wiener filtering, MMSE STSA, MMSE LSA, OM LSA, etc.) [1], [2], [6], [9]–[11]. The resulting speech spectrum is then estimated by applying the spectral gain to the noisy spectrum

$$\hat{S}(p,k) = G(p,k)X(p,k). \tag{5}$$

III. SNR Analysis Tool

In order to evaluate the behavior of speech enhancement techniques, we propose to use an approach described by Renevey and Drygajlo [12]. The basic principle is to consider the *a priori* SNR as a function of the *a posteriori* SNR in order to analyze the behavior of the features defined by the 2-tuple $(\mathrm{SNR}_{\mathrm{post}}, \mathrm{SNR}_{\mathrm{prio}})$.

In the additive model, the amplitude of the noisy signal can be expressed as

$$\begin{aligned}&|X(p,k)|\\ &= \sqrt{|S(p,k)|^2 + |N(p,k)|^2 + 2\,|S(p,k)|\,|N(p,k)|\cos\alpha(p,k)}\end{aligned} \tag{6}$$

where $\alpha(p,k)$ is the phase difference between $S(p,k)$ and $N(p,k)$. Assuming the knowledge of the clean speech and the noise, the local *a posteriori* and *a priori* SNRs, can be defined by

$$\mathrm{SNR}_{\mathrm{post}}^{\mathrm{local}}(p,k) = \frac{|X(p,k)|^2}{|N(p,k)|^2} \tag{7}$$

and

$$\mathrm{SNR}_{\mathrm{prio}}^{\mathrm{local}}(p,k) = \frac{|S(p,k)|^2}{|N(p,k)|^2}. \tag{8}$$

By replacing $|X(p,k)|$ in (7) by its expression (6) and using (8), we get

$$\begin{aligned}\mathrm{SNR}_{\mathrm{post}}^{\mathrm{local}}(p,k) &= 1 + \mathrm{SNR}_{\mathrm{prio}}^{\mathrm{local}}(p,k)\\ &\quad + 2\sqrt{\mathrm{SNR}_{\mathrm{prio}}^{\mathrm{local}}(p,k)}\cos\alpha(p,k).\end{aligned} \tag{9}$$

Note that this relationship depends on $\alpha(p,k)$ which is an uncontrolled parameter in classic speech enhancement techniques. For example, in the derivation of the classic Wiener filter, the $\mathrm{SNR}_{\mathrm{post}}(p,k)$ is assumed to be equal to $1 + \mathrm{SNR}_{\mathrm{prio}}(p,k)$

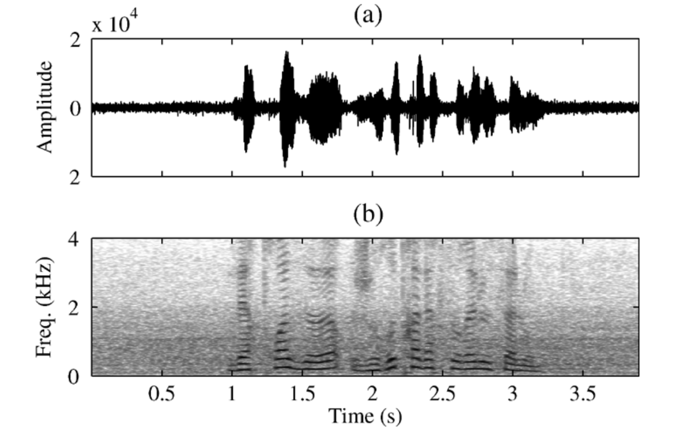

Fig. 1. (a) Waveform and (b) spectrum of the French sentence “Vers trois heures je re-traverserai le salon.” corrupted by car noise at 12-dB global SNR.

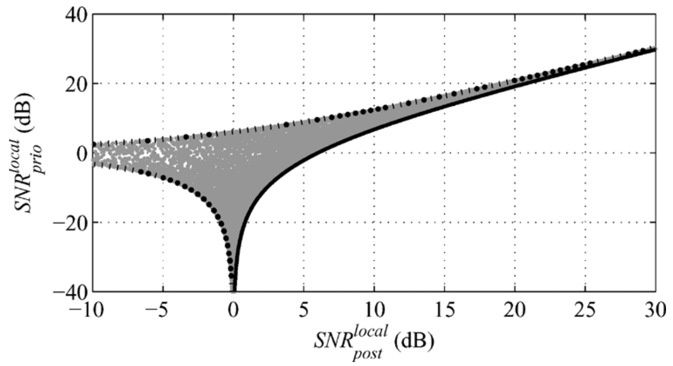

Fig. 2. $\mathrm{SNR}_{\mathrm{prio}}^{\mathrm{local}}$ versus $\mathrm{SNR}_{\mathrm{post}}^{\mathrm{local}}$ assuming the knowledge of clean speech and noise amplitudes. The two lines illustrate (9) when $\alpha(p,k)=0$ (solid line) and $\alpha(p,k)=\pi$ (dashed line).

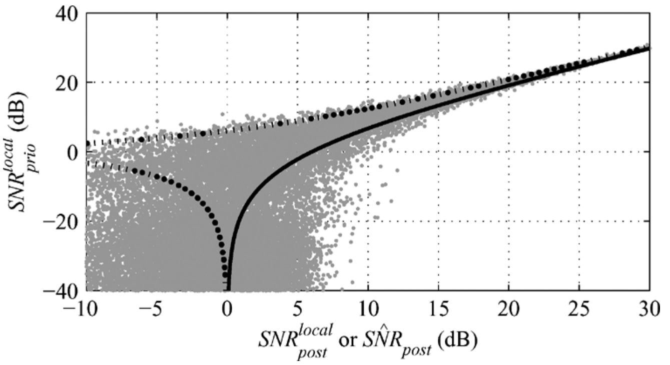

Fig. 3. $\mathrm{SNR}_{\mathrm{prio}}^{\mathrm{local}}$ versus $\mathrm{SNR}_{\mathrm{post}}^{\mathrm{local}}$ assuming the knowledge of clean speech amplitude but the noise PSD being estimated. The two lines illustrate (9) when $\alpha(p,k)=0$ (solid line) and $\alpha(p,k)=\pi$ (dashed line).

which corresponds to a constant phase difference $\alpha(p,k)=\pi/2$ (i.e., noise and clean speech are supposed to add in quadrature).

In the following, the discussion will be illustrated using a sentence corrupted by car noise at 12-dB global SNR, but it can be generalized to other noise types and SNR conditions. The waveform and spectrum of this signal are shown in Fig. 1(a) and (b), respectively. The relationship expressed by (9) is illustrated in Fig. 2. It presents the *a priori* SNR versus the *a posteriori* SNR in the ideal case where the clean speech and noise amplitudes are known. The features lie between two curves, the solid one (resp. dashed) corresponds to the limit case where $\alpha(p,k)=0$ (resp. π), i.e., noise and clean speech spectral components add in phase (resp. phase opposition). These two limits define an area where the feature distribution depends on the true phase difference $\alpha(p,k)$. Note that since only the amplitudes of the signals are used to obtain the SNRs involved in the spectral gain computation, estimation errors inherent to the speech enhancement method cannot be avoided even knowing the clean speech.

Fig. 3 illustrates the case where an estimation of the noise PSD is used in (7) and (8) instead of the local noise but still assuming the knowledge of the clean speech amplitude. In that case, the $\mathrm{SNR}_{\mathrm{post}}^{\mathrm{local}}$ corresponds to $\hat{\mathrm{SNR}}_{\mathrm{post}}$ of (1). The noise PSD estimation errors lead to an important feature dispersion outside of the boundary for low SNR values and slightly decrease the quality of the enhanced speech. Given a noise PSD estimation, this is the case leading to the better SNR estimates. It will then be used as a reference in the next sections.

IV. Decision-Directed Approach

A. Principle of the Decision-Directed Algorithm

In the sequel, we use a classic noise estimation based on voice activity detection [1] (in contrast with continuous estimations [7], [8]). Using the obtained noise PSD, the *a posteriori* and *a priori* SNRs are computed as follows:

$$\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k)=\frac{|X(p,k)|^2}{\hat{\gamma}_n(p,k)} \tag{10}$$

and

$$\begin{aligned}\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)=\beta&\frac{\left|\hat{S}(p-1,k)\right|^2}{\hat{\gamma}_n(p,k)}\\&+(1-\beta)\mathrm{P}\left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k)-1\right]\end{aligned} \tag{11}$$

where P[.] denotes the half-wave rectification, and $\hat{S}(p-1,k)$ is the estimated speech spectrum at previous frame. This *a priori* SNR estimator corresponds to the so-called decision-directed approach [2], [3] whose behavior is controlled by the parameter β (typically $\beta=0.98$). Without loss of generality, in the following the chosen spectral gain [function g in (4)] is the Wiener filter, and then

$$G_{\mathrm{DD}}(p,k)=\frac{\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)}{1+\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)}. \tag{12}$$

The approach defined by (10)–(12) is called the DD algorithm.

B. Analysis of the DD Algorithm

We can emphasize two effects of the DD algorithm which have been interpreted by Cappé in [3].

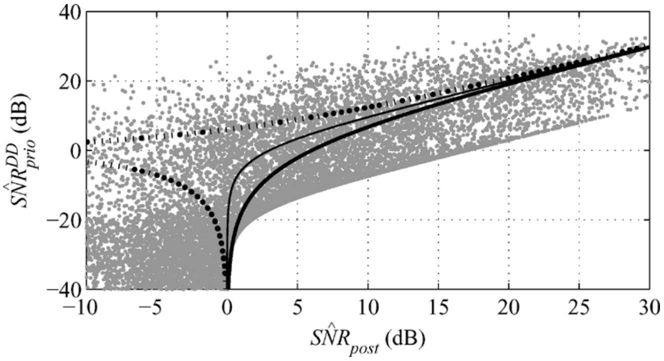

Fig. 4. $\hat{\mathrm{SNR}}^{\mathrm{DD}}_{\mathrm{prio}}$ versus $\hat{\mathrm{SNR}}_{\mathrm{post}}$ for the DD approach. The three lines illustrate (9) when $\alpha(p,k) = 0$ (bold solid line), $\alpha(p,k) = \pi$ (dashed line), and $\alpha(p,k) = \pi/2$ (thin solid line).

- When the *instantaneous* SNR is much larger than 0 dB, $\hat{\mathrm{SNR}}_{\mathrm{prio}}(p,k)$ corresponds to a frame delayed version of the *instantaneous* SNR.
- When the *instantaneous* SNR is lower or close to 0 dB, $\hat{\mathrm{SNR}}_{\mathrm{prio}}(p,k)$ corresponds to a highly smoothed and delayed version of the *instantaneous* SNR. Thus, the variance of the *a priori* SNR is reduced compared to the *instantaneous* SNR. The direct consequence for the enhanced speech is the reduction of the musical noise effect.

The delay inherent to the DD algorithm is a drawback especially in the speech transients, e.g., speech onset and offset. Furthermore, this delay introduces a bias in gain estimation which limits noise reduction performance and generates an annoying reverberation effect.

In order to describe the behavior of the DD approach, the 2-tuple ($\hat{\mathrm{SNR}}_{\mathrm{post}}, \hat{\mathrm{SNR}}^{\mathrm{DD}}_{\mathrm{prio}}$) is represented in Fig. 4, where the *a posteriori* and *a priori* SNRs are estimated using (10) and (11), respectively. To analyze this figure, the reference is the case when SNRs are computed using known clean speech amplitude and estimated noise PSD (cf. Fig. 3). In Fig. 4 a large part of the *a priori* SNR features (approximately 60% in this case) is underestimated which illustrates the effect of the DD bias on SNR estimation.

If we consider the case where a speech component appears abruptly at frame p, assuming that the *a priori* SNR is zero at frame $p-1$, then for the current frame we have

$$\hat{\mathrm{SNR}}^{\mathrm{DD}}_{\mathrm{prio}}(p,k) = (1-\beta)\mathrm{P}\left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) - 1\right]. \quad (13)$$

Actually, the estimated *a priori* SNR will be a version of the *instantaneous* SNR attenuated by $(1-\beta)$. A typical value $\beta = 0.98$ leads to an attenuation of almost 17 dB. Note that if $\alpha(p,k) = \pi/2$, (9) becomes

$$\mathrm{SNR}^{\mathrm{local}}_{\mathrm{prio}}(p,k) = \mathrm{SNR}^{\mathrm{local}}_{\mathrm{post}}(p,k) - 1 = \mathrm{SNR}^{\mathrm{local}}_{\mathrm{inst}}(p,k). \quad (14)$$

This relationship is illustrated in Fig. 4 by the thin solid line. Thus, the attenuation introduced by $1-\beta$ in (13) is materialized by a high concentration of features around a shifted version (by -17 dB) of this thin line curve. This offset corresponds to the maximum bias and it is consistent with the degradation introduced by the DD approach during speech onsets and more generally when speech amplitude increases rapidly. Note that if β increases, the bias increases too, further reducing the musical noise but introducing a larger underestimation of the *a priori* SNR.

We can also observe in Fig. 4 that some *a priori* SNR features are overestimated. This case occurs when a speech component disappears abruptly, i.e., $\mathrm{P}[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) - 1] = 0$ leading to

$$\hat{\mathrm{SNR}}^{\mathrm{DD}}_{\mathrm{prio}}(p,k) = \beta \frac{\left|\hat{S}(p-1,k)\right|^2}{\hat{\gamma}_n(p,k)} \quad (15)$$

whereas a null value would be the best estimate. This overestimation is related to the speech spectrum of the previous frame. The reverberation effect characteristic of the DD approach is explained by both underestimation and overestimation of the *a priori* SNR features.

C. Comparison Between A Posteriori and A Priori SNRs

It is interesting to underline the behavior of the *a posteriori* and *a priori* SNR estimators. It is well known that using only the *a posteriori* SNR to enhance the noisy speech results in a very high amount of musical noise, leading to a poor signal quality. However, this technique leads to the lowest degradation level for the speech components themselves. The *a priori* SNR, estimated in the DD approach, is widely used instead of the *a posteriori* SNR because the musical noise is reduced to an acceptable level. However, this estimated SNR is biased and then the performance is reduced during speech activity. From a subjective point of view, this bias is perceived as a reverberation effect.

In order to measure the performance of SNR estimators, it is useful to compare the estimated SNR values to the true (local) ones as shown in Fig. 5, where the estimated SNRs are displayed versus the true SNRs in (7) and (8). The SNRs are plotted for 50 frames of speech activity to focus the analysis on the behavior of the SNR estimators for speech components.

Fig. 5(a) illustrates the *a posteriori* SNR estimated in the way proposed in (10) and Fig. 5(b) the *a priori* SNR estimated using the DD approach given by (11). In these two cases, the bold line corresponds to a perfect SNR estimator ($\hat{\mathrm{SNR}} = \mathrm{SNR}^{\mathrm{local}}$) that can be used as a reference to evaluate the performance of the real estimators. It is obvious that the features corresponding to the *a posteriori* SNR estimator are closer to the reference bold line and less dispersed than the *a priori* SNR estimator ones.

The dispersion observed for the two cases (a) and (b) of Fig. 5 can be characterized by the correlation coefficient which can be computed as

$$\rho = \frac{\mathrm{E}\left[\left(\hat{\mathrm{SNR}} - \mathrm{E}[\hat{\mathrm{SNR}}]\right)\left(\mathrm{SNR}^{\mathrm{local}} - \mathrm{E}[\mathrm{SNR}^{\mathrm{local}}]\right)\right]}{\sqrt{\mathrm{E}\left[\left(\hat{\mathrm{SNR}} - \mathrm{E}[\hat{\mathrm{SNR}}]\right)^2\right]\mathrm{E}\left[\left(\mathrm{SNR}^{\mathrm{local}} - \mathrm{E}[\mathrm{SNR}^{\mathrm{local}}]\right)^2\right]}}. \quad (16)$$

For typical cases depicted in Fig. 5, we obtain $\rho_{\mathrm{post}} = 0.79$ and $\rho_{\mathrm{prio}} = 0.23$ which is consistent with the observed feature dispersion for the two cases (a) and (b) of Fig. 5, a smaller correlation coefficient leading to a greater dispersion. When generalizing to a wider range of noise types and SNR levels, it was

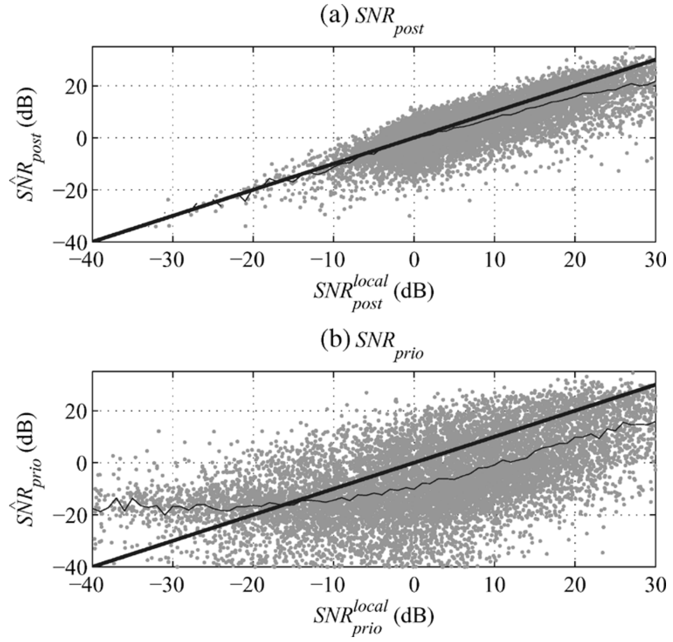

Fig. 5. Estimated SNRs versus true SNRs (i.e., local SNRs) in case of (a) *a posteriori* SNR and (b) *a priori* SNR. The bold line represents a perfect estimator, and the thin line represents the mean of the estimated SNR versus the true SNR.

observed that ρ_{prio} and ρ_{post} are related by the following equation:

$$\rho_{\mathrm{prio}} \approx \rho_{\mathrm{post}} - 0,5. \tag{17}$$

In Fig. 5(a) and (b), the thin line represents the mean of the estimated SNR knowing the true SNR and is theoretically obtained as follows:

$$\mathrm{E}[\hat{\mathrm{SNR}}|\mathrm{SNR}^{\mathrm{local}}] = \int \hat{\mathrm{snr}}\ \mathrm{p}(\hat{\mathrm{snr}}|\mathrm{SNR}^{\mathrm{local}})\mathrm{d}\hat{\mathrm{snr}} \tag{18}$$

where p is the probability density function. The mean of the estimated SNR is closer to the perfect estimator for the *a posteriori* SNR estimator. It is slightly underestimated for high SNR, whereas for the *a priori* SNR the underestimation is large for SNR greater than -17 dB. However, since the dispersion is high for the *a priori* SNR features, even if the mean is largely underestimated, the case where SNR features are overestimated exists. Furthermore, the *a priori* SNR is overestimated for SNR smaller than -17 dB. Finally, these results confirm that the *a posteriori* SNR estimator is more reliable than the *a priori* SNR estimator for speech components.

V. Two-Step Noise Reduction Technique

A. Principle of the TSNR Technique

In order to enhance the performance of the noise reduction process, we propose to estimate the *a priori* SNR in a two-step procedure. The DD algorithm introduces a frame delay when the parameter β is close to one. Consequently, the spectral gain computed at current frame p matches the previous frame $p-1$. Based on this fact, we propose to compute the spectral gain for the next frame $p+1$ using the DD approach and to apply it to the current frame because of the frame delay. This leads to an algorithm in two steps.

In the first step, using the DD algorithm, we compute the spectral gain $G_{\mathrm{DD}}(p,k)$ as described in (12). In the second step, this gain is used to estimate the *a priori* SNR at frame $p+1$

$$\begin{aligned}\hat{\mathrm{SNR}}^{\mathrm{TNSR}}_{\mathrm{prio}}(p,k) &= \hat{\mathrm{SNR}}^{\mathrm{DD}}_{\mathrm{prio}}(p+1,k)\\ &= \beta' \frac{|G_{\mathrm{DD}}(p,k)X(p,k)|^2}{\hat{\gamma}_n(p,k)} + (1-\beta')\mathrm{P}\\ &\quad \left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p+1,k)-1\right]\end{aligned} \tag{19}$$

where β' plays the same role as β but can have a different value. Note that to compute $\hat{\mathrm{SNR}}_{\mathrm{post}}(p+1,k)$ we need the knowledge of the future frame $X(p+1,k)$ which introduces an additional processing delay and may be incompatible with the desired application. Thus, we propose to choose $\beta' = 1$, in this case the previous estimator of (19) degenerates into the following particular case:

$$\hat{\mathrm{SNR}}^{\mathrm{TNSR}}_{\mathrm{prio}}(p,k) = \frac{|G_{\mathrm{DD}}(p,k)X(p,k)|^2}{\hat{\gamma}_n(p,k)}. \tag{20}$$

This avoids introducing an additional processing delay since the term using the future is not required. Furthermore, as $\beta' = 1$, the musical noise level will be reduced to the lowest level allowed by the DD approach. The choice of $\beta' = 1$ is valid only for the second step in order to refine the first step estimation: actually β is set to a typical value of 0.98 for the first step.

Finally, we compute the spectral gain

$$G_{\mathrm{TNSR}}(p,k) = h\left(\hat{\mathrm{SNR}}^{\mathrm{TNSR}}_{\mathrm{prio}}(p,k), \hat{\mathrm{SNR}}_{\mathrm{post}}(p,k)\right) \tag{21}$$

which is used to enhance the noisy speech

$$\hat{S}(p,k) = G_{\mathrm{TNSR}}(p,k)X(p,k). \tag{22}$$

Note that h may be different from the function g defined in (4). However, without loss of generality, in the following the chosen spectral gain is the Wiener filter too, and then

$$G_{\mathrm{TNSR}}(p,k) = \frac{\hat{\mathrm{SNR}}^{\mathrm{TNSR}}_{\mathrm{prio}}(p,k)}{1+\hat{\mathrm{SNR}}^{\mathrm{TNSR}}_{\mathrm{prio}}(p,k)}. \tag{23}$$

This algorithm in two steps defined by (10), (11), (20), and (23) is called the TSNR technique.

B. Theoretical Analysis of the TSNR Technique

The noisy signal described in Section III has been processed by DD and TSNR algorithms. The typical behaviors of these algorithms are illustrated in Fig. 6 where the time varying SNRs

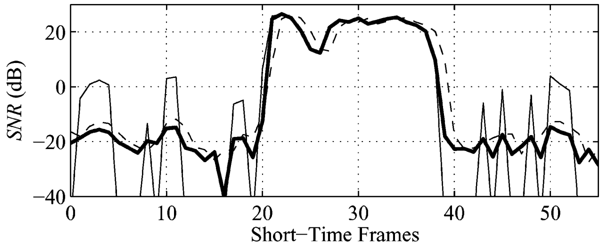

Fig. 6. SNR evolution over short-time frames ($f = 467$ Hz). Thin solid line: *instantaneous* SNR; dashed line: *a priori* SNR for the DD algorithm; bold solid line: *a priori* SNR for the TSNR algorithm.

at frequency 467 Hz are displayed. The first 20 frames and the last 17 contain only car noise and the frames in between contain noisy speech ($\mathrm{SNR} = 12$ dB) including speech onset and offset. The thin solid line represents the time varying instantaneous SNR. The dashed line and the bold solid one represent the *a priori* SNR evolutions for the DD algorithm and for the TSNR algorithm, respectively. From Fig. 6, the behavior of the TSNR algorithm can be described as follows.

- When the *instantaneous* SNR is much larger than 0 dB, $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}(p,k)$ follows the *instantaneous* SNR without delay contrary to $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)$. Furthermore, when $\hat{\mathrm{SNR}}_{\mathrm{inst}}(p,k)$ increases or decreases (speech onset or offset), the response of $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}(p,k)$ is also instantaneous while that of $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)$ is delayed.
- When the *instantaneous* SNR is lower than or close to 0 dB, the $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}(p,k)$ is further reduced compared to $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k)$. Furthermore, it appears that the second step helps in reducing the delay introduced by the smoothing effect even when the SNR is small, while keeping the desired smoothing effect. This behavior is consistent with the fact that $\beta' = 1$ in the second step (20) which is a decision-directed estimator too, so by increasing β' the residual musical noise is reduced to the lowest level allowed by the DD approach.

To summarize, the TSNR algorithm improves the noise reduction performance since the gain matches to the current frame whatever the SNR. The main advantages of this approach are the ability to preserve speech onsets and offsets, and to successfully remove the annoying reverberation effect typical of the DD approach. Note that in practice this reverberation effect can be reduced by increasing the overlap between successive frames but cannot be suppressed whereas the TSNR approach makes it possible with a typical overlap of 50%.

An analysis of the TSNR algorithm using the 2-tuple $(\hat{\mathrm{SNR}}_{\mathrm{post}}, \hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}})$ representation is depicted in Fig. 7. It is possible to distinguish two asymptotical behaviors corresponding to high point density in the feature space.

The case corresponding to the lower limit of the features occurs when no speech is present in the previous frame $p-1$ leading to $\hat{S}(p-1,k) = 0$. Then, at frame p, the DD approach gives the following estimation for the *a priori* SNR:

$$\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k) = (1-\beta)\mathrm{P}\left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) - 1\right] \tag{24}$$

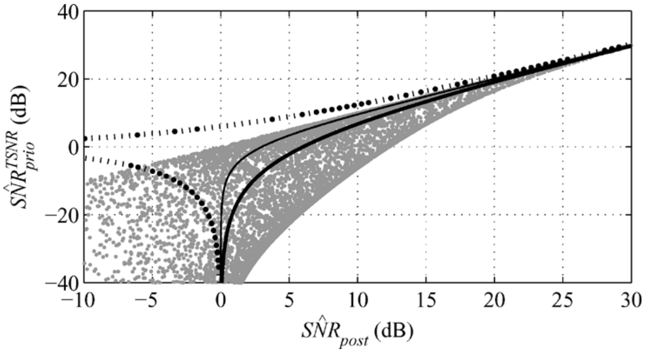

Fig. 7. $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}$ versus $\hat{\mathrm{SNR}}_{\mathrm{post}}$ for the TSNR approach. The three lines illustrate (9) when $\alpha(p,k) = 0$ (bold solid line), $\alpha(p,k) = \pi$ (dashed line), and $\alpha(p,k) = \pi/2$ (thin solid line).

which introduces an attenuation of almost 17 dB if $\beta = 0.98$. When refining the *a priori* SNR estimation by the second step according to (20) and using (12) and (10), the TSNR approach leads to

$$\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}(p,k) = \left(\frac{(1-\beta)\mathrm{P}\left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) - 1\right]}{1 + (1-\beta)\mathrm{P}\left[\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) - 1\right]}\right)^2 \hat{\mathrm{SNR}}_{\mathrm{post}}(p,k). \tag{25}$$

By searching the intersection between the curves defined by (24) and (25), we show that if

$$\hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) > \frac{1}{2\beta}\left(1 + 2\beta + \sqrt{\frac{1+3\beta}{1-\beta}}\right) \tag{26}$$

then the TSNR approach delivers a greater SNR than the DD one. Classically, $\beta = 0.98$, and this threshold is almost equal to 9.4 dB. Consequently, if a signal component appears abruptly at frame p, thus increasing the *a posteriori* SNR, the estimated *a priori* SNR tends to the *a posteriori* SNR suppressing the bias introduced by the DD approach. This bias decreases when the *a posteriori* SNR increases. However, if speech is absent at frame p too, keeping the *a posteriori* SNR to a low level, the estimated *a priori* SNR becomes lower than for the DD approach further limiting the musical noise.

The case corresponding to the upper limit of the features of Fig. 7 essentially occurs when the *a priori* SNR is high (overestimated by DD approach or not) at frame $p-1$ and becomes low at frame p, i.e., when the spectral speech component decays rapidly. In that case, we can derive from (11) the following approximation [3]:

$$\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{DD}}(p,k) \approx \beta\hat{\mathrm{SNR}}_{\mathrm{inst}}(p-1,k). \tag{27}$$

So, the spectral gain obtained after the first step can be approximated by

$$G_{\mathrm{DD}}(p,k) \approx \frac{\beta\hat{\mathrm{SNR}}_{\mathrm{inst}}(p-1,k)}{1+\beta\hat{\mathrm{SNR}}_{\mathrm{inst}}(p-1,k)}. \tag{28}$$

Furthermore, by considering that $\hat{\mathrm{SNR}}_{\mathrm{inst}}(p-1,k) \gg 1$ and that β is very close to 1, (28) reduces to $G_{\mathrm{DD}}(p,k) \approx 1$. If we introduce this approximation in (20), this leads to

$$\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{TNSR}}(p,k) \approx \hat{\mathrm{SNR}}_{\mathrm{post}}(p,k) \approx \hat{\mathrm{SNR}}_{\mathrm{inst}}(p,k) \quad (29)$$

which explains that the shape of the upper limit is a straight line. This refinement suppresses the *a priori* SNR overestimation.

As a conclusion, the TSNR approach has the ability to preserve speech onsets and offsets and is able to suppress the reverberation effect typical of the DD approach. For high SNR, the *a priori* SNR underestimation which is due to the delay introduced by the DD approach is suppressed while for low SNR the underestimation is preserved in order to achieve the musical noise suppression. The *a priori* SNR overestimation is also suppressed.

VI. Speech Harmonic Regeneration

The output signal $\hat{S}(p,k)$, or $\hat{s}(t)$ in the time domain, obtained by the TSNR technique presented in the previous section still suffers from distortions. This is inherent to the estimation errors introduced by the noise spectrum estimation since it is very difficult to get reliable instantaneous estimates in single channel noise reduction techniques. Since 80% of the pronounced sounds are voiced in average, the distortions generally turn out to be harmonic distortion. Indeed, some harmonics are considered as noise-only components and are suppressed. We propose to take advantage of the harmonic structure of voiced speech to prevent this distortion. For that purpose, we propose to process the distorted signal to create a fully harmonic signal where all the missing harmonics are regenerated. This signal will then be used to compute a spectral gain able to preserve the speech harmonics. This will be called the speech harmonic regeneration step and can be used to improve the results of any noise reduction technique and not only the TSNR one.

A. Principle of Harmonic Regeneration

A simple and efficient way to restore speech harmonics consists of applying a nonlinear function NL (e.g., absolute value, minimum, or maximum relative to a threshold, etc.) to the time signal enhanced in a first procedure with a classic noise reduction technique. Then, the artificially restored signal $s_{\mathrm{harmo}}(t)$ is obtained by

$$s_{\mathrm{harmo}}(t) = NL\left(\hat{s}(t)\right). \quad (30)$$

Note that the restored harmonics of $s_{\mathrm{harmo}}(t)$ are created at the same positions as the clean speech ones. This very interesting and important characteristic is implicitly ensured because a nonlinearity in the time domain is used to restore the harmonics. For illustration, Fig. 8 shows the typical effect of the nonlinearity and illustrates its interest. Fig. 8(a) represents a reference frame of voiced clean speech. Fig. 8(b) represents the same frame after being corrupted by noise and processed by the TSNR algorithm presented in Section V. It appears clearly that some harmonics have been completely suppressed or severely degraded. Fig. 8(c) represents the artificially restored frame obtained using (30) where the nonlinearity (half wave rectification, i.e., the maximum relative to 0, has been used in this example) applied to the signal $\hat{s}(t)$ has restored the suppressed or degraded harmonics at the same positions as in clean speech. However, the harmonic amplitudes of this artificial signal are biased compared to clean speech. As a consequence, this signal $s_{\mathrm{harmo}}(t)$ cannot be used directly as clean speech estimation. Nevertheless, it contains a very useful information that can be exploited to refine the *a priori* SNR

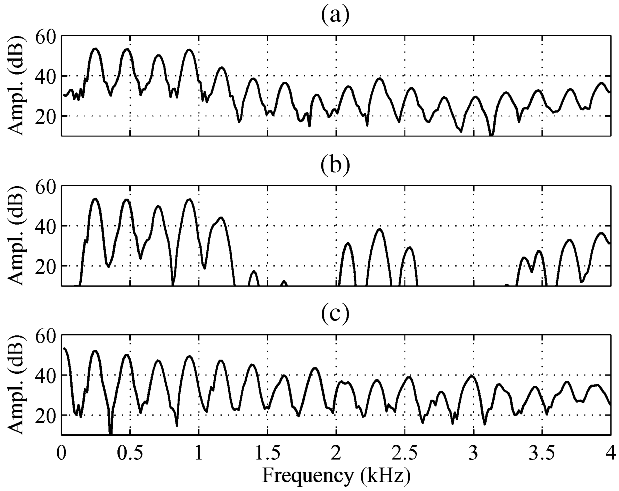

Fig. 8. Effect of the nonlinearity on a voiced frame. (a) Clean speech spectrum. (b) Enhanced speech spectrum using TSNR technique. (c) Artificially restored speech spectrum after harmonic regeneration.

$$\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{HRNR}}(p,k) = \frac{\rho(p,k)\left|\hat{S}(p,k)\right|^2 + (1-\rho(p,k))\left|S_{\mathrm{harmo}}(p,k)\right|^2}{\hat{\gamma}_n(p,k)}. \quad (31)$$

The $\rho(p,k)$ parameter is used to control the mixing level of $|\hat{S}(p,k)|^2$ and $|S_{\mathrm{harmo}}(p,k)|^2$ $(0 \le \rho(p,k) \le 1)$. This mixing is necessary because the nonlinear function is able to restore harmonics at the desired frequencies, but with biased amplitudes. Then the behavior of this parameter should be:

- when the estimation of $\hat{S}(p,k)$ provided by the TSNR algorithm (for example) is reliable, the harmonic regeneration process is not needed and $\rho(p,k)$ should be equal to 1;
- when the estimation of $\hat{S}(p,k)$ provided by the TSNR algorithm is unreliable, the harmonic regeneration process is required to correct the estimation and $\rho(p,k)$ should be equal to 0 (or any other constant value depending on the chosen nonlinear function).

We propose to choose $\rho(p,k) = G_{\mathrm{TNSR}}(p,k)$ to match this behavior. The $\rho(p,k)$ parameter can also be chosen constant to realize a compromise between the two estimators $\hat{S}(p,k)$ and $S_{\mathrm{harmo}}(p,k)$.

The refined *a priori* SNR, $\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{HRNR}}(p,k)$, is then used to compute a new spectral gain which will be able to preserve the harmonics of the speech signal

$$G_{\mathrm{HRNR}}(p,k) = v\left(\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{HRNR}}(p,k), \hat{\mathrm{SNR}}_{\mathrm{post}}(p,k)\right). \quad (32)$$

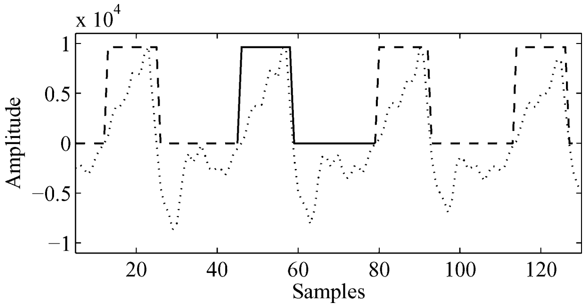

Fig. 9. Voiced speech frame $\hat{s}(t)$ (dotted line) and associated scaled $p(\hat{s}(t))$ signal (dashed line). Repeated elementary waveform (solid line).

The function v can be chosen among the different gain functions proposed in the literature (e.g., amplitude or power spectral subtraction, Wiener filtering, etc.) [1], [2], [6], [9]–[11]. Without loss of generality, in the following, the chosen spectral gain is the Wiener filter, and then

$$G_{\mathrm{HRNR}}(p,k)=\frac{\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{HRNR}}(p,k)}{1+\hat{\mathrm{SNR}}_{\mathrm{prio}}^{\mathrm{HRNR}}(p,k)}. \tag{33}$$

Finally, the resulting speech spectrum is estimated as follows:

$$\hat{S}(p,k)=G_{\mathrm{HRNR}}(p,k)X(p,k). \tag{34}$$

This approach, defined by (30), (31), and (33), which has the ability to preserve the harmonics suppressed by classic algorithms and thus avoids distortions, is called the harmonic regeneration noise reduction (HRNR) technique.

B. Theoretical Analysis of Harmonic Regeneration

To analyze the harmonic regeneration step, we will focus on a particular nonlinearity, without loss of generality, the half wave rectification. Replacing the nonlinear function NL by the Max function in (30), it follows that

$$s_{\mathrm{harmo}}(t)=\mathrm{Max}\left(\hat{s}(t),0\right)=\hat{s}(t)p\left(\hat{s}(t)\right) \tag{35}$$

where p is defined as

$$p(u)=\begin{cases}1, & \text{if } u>0\\ 0, & \text{if } u<0.\end{cases} \tag{36}$$

Fig. 9 represents a frame of the voiced speech signal $\hat{s}(t)$ (dotted line) and the corresponding $p(\hat{s}(t))$ signal (dashed line). Note that this signal is scaled to make the figure clearer. It can be observed that the signal $p(\hat{s}(t))$ amounts to a repetition of an elementary waveform (solid line) with periodicity T, corresponding to the voiced speech pitch period. Assuming the quasistationarity of speech over a frame duration, the Fourier transform (FT) of $p(\hat{s}(t))$ comes down to a sampled version (by $1/T$ steps) of the elementary waveform's FT

$$\mathrm{FT}\left(p\left(\hat{s}(t)\right)\right)=\frac{1}{T}\sum_{m=-\infty}^{\infty}R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{37}$$

where δ corresponds to the Dirac distribution, f denotes the continuous frequency, and $R(m/T)$ is the FT of the elementary waveform taken at discrete frequency m/T. Note that the sampling frequency coincides with the harmonic positions of the elementary waveform. Finally, using (35), the FT of $s_{\mathrm{harmo}}(t)$ can be written as

$$\mathrm{FT}\left(s_{\mathrm{harmo}}(t)\right)=\mathrm{FT}\left(\hat{s}(t)\right)*\frac{e^{-j\theta}}{T}\sum_{m=-\infty}^{\infty}R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{38}$$

where θ is the phase at origin. Thus, the spectrum of the restored signal, $s_{\mathrm{harmo}}(t)$, is the convolution between the spectrum of $\hat{s}(t)$, signal enhanced by the TSNR as in Fig. 8(b), and an harmonic comb. This comb has the same fundamental frequency as the voiced speech signal $\hat{s}(t)$ which explains the phenomenon of harmonic regeneration. The main advantage of this method is its simplicity to restore speech harmonics at desired positions. Furthermore, the envelope of $\mathrm{FT}(p(\hat{s}(t)))$, symmetric about $m=0$, is rapidly decreasing when $|m|$ increases, thus a missing harmonic is regenerated only using the information of the few neighboring harmonics. Of course, because of this behavior, the harmonic regeneration process will be less efficient if too many harmonics are missing, e.g., signal with too small input SNR).

It is also important to investigate the behavior of the harmonic regeneration process for unvoiced speech. Let us consider a hybrid signal where the lower part of the spectrum is voiced and the upper part unvoiced. The FT of $p(\hat{s}(t))$ (37) will still be an harmonic comb, its fundamental frequency being imposed by the voiced lower part of the spectrum. Then the spectrum of the resulting signal $\mathrm{FT}(s_{\mathrm{harmo}}(t))$ will be the result of (38) exactly as in voiced only speech case. However, since the envelope of the harmonic comb is rapidly decreasing, each frequency bin is obtained using only its corresponding neighboring area in the spectrum of $\hat{s}(t)$. Then, the unvoiced spectrum part will lead to an unvoiced restored spectrum since the harmonics of the spectrum of $\hat{s}(t)$ will not be used to restore the unvoiced part.

Now, let us consider the case where the full band of speech is unvoiced. The FT of $p(\hat{s}(t))$ (37) is obviously not an harmonic comb, it will be an undetermined spectrum. However, the convolution in (38) between the unvoiced spectrum and this undetermined spectrum will automatically lead to an unvoiced spectrum. Thus, in that case too, the unvoiced parts of speech will not be degraded by the harmonic regeneration process.

This behavior for unvoiced speech components ensures that unvoiced speech parts are not degraded by the harmonic regeneration process.

C. Illustration of HRNR Behavior

The principle and an analysis of the HRNR technique have been proposed in the previous sections. We propose to illustrate its behavior and performance in a typical case of noisy speech. Fig. 10 shows four spectrograms, Fig. 10(a) represents the noisy speech in the context described in Section III (car noise at 12-dB global SNR), Fig. 10(b) and (c) shows the enhanced noisy speech using the TSNR and HRNR techniques, respectively. Fig. 10(d) represents the clean speech and is, therefore, the reference to compare the results obtained by TSNR and HRNR approaches. Note that no threshold is used to constraint the noise reduction filter of each algorithm to make the

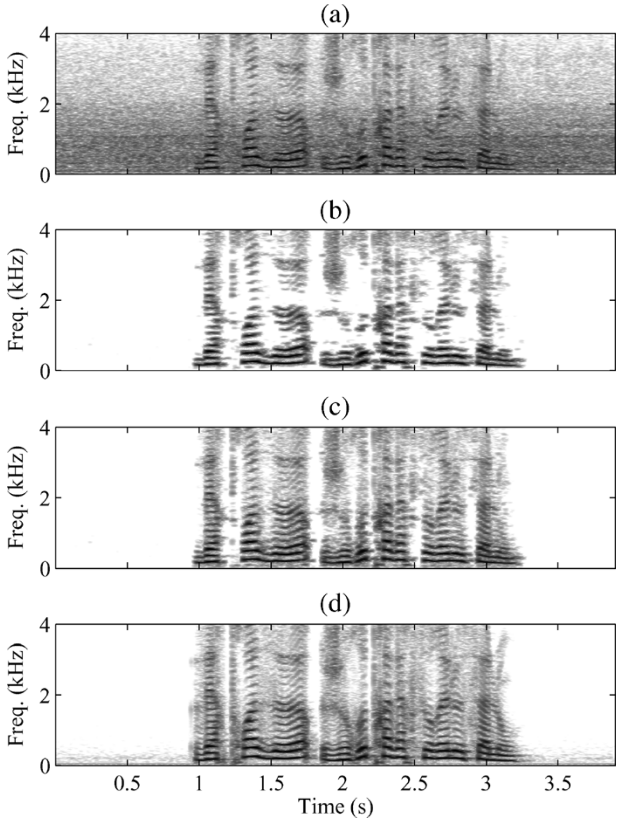

Fig. 10. Speech spectrograms. (a) Noisy speech corrupted by car noise at 12-dB SNR. (b) Noisy speech enhanced by TSNR technique. (c) Noisy speech enhanced by HRNR technique. (d) Clean speech.

spectrograms clearer. By comparing cases (b)–(d) in Fig. 10, it appears that many harmonics are preserved using HRNR technique, whereas they are suppressed when using TSNR. So, this example shows that taking into account the voiced characteristic of speech can be used to enhance harmonics completely degraded by noise.

VII. Results

The output of the TSNR technique is used as an input of the HRNR technique. Hence, the comparison of results obtained for both techniques will give the improvement brought by the harmonic regeneration process alone. The TSNR technique will then be used as the reference. The sampling frequency of the processed signals is 8 kHz. Accordingly, the following parameters have been chosen: frame size $L = 256$ (32 ms), windows overlap 50%, the size of the fast Fourier transform (FFT) $N_{\mathrm{FFT}} = 512$. Recall that the spectral gain used for both algorithms [g in (4), h in (21), and v in (32)] is the Wiener filter [cf. (12), (23), and (34)]. In the TSNR technique, the parameters are $\beta = 0.98$ and $\beta' = 1$. In the HRNR technique, the chosen nonlinear function is the half wave rectification (cf. (35)), and the rule retained for the mixing parameter of (31) is $\rho(p,k) = G_{\mathrm{TNSR}}(p,k)$.

A. Objective Results

To measure the performance of the TSNR and HRNR techniques, we chose the cepstral distance (CD) [13] as it is a degradation measure correlated with subjective tests results. It is usually admitted that the distortion is not audible if the CD is below 0.5. An example is given in Fig. 11 based on the noisy speech of Fig. 10(a). This figure shows the time variations of the CD between clean speech and speech enhanced by TSNR technique, Fig. 11(b), and speech enhanced by HRNR technique, Fig. 11(c), respectively. The clean speech is displayed in Fig. 11(a) to ease the interpretation of the CDs. The CD for HRNR technique is smaller than for TSNR technique, therefore the HRNR technique introduces less distortions than the TSNR resulting in a better quality of the enhanced speech. Note that in Fig. 11(b) and (c), high peaks are located in low energy zones [cf. Fig. 11(a)] which are of low perceptual importance.

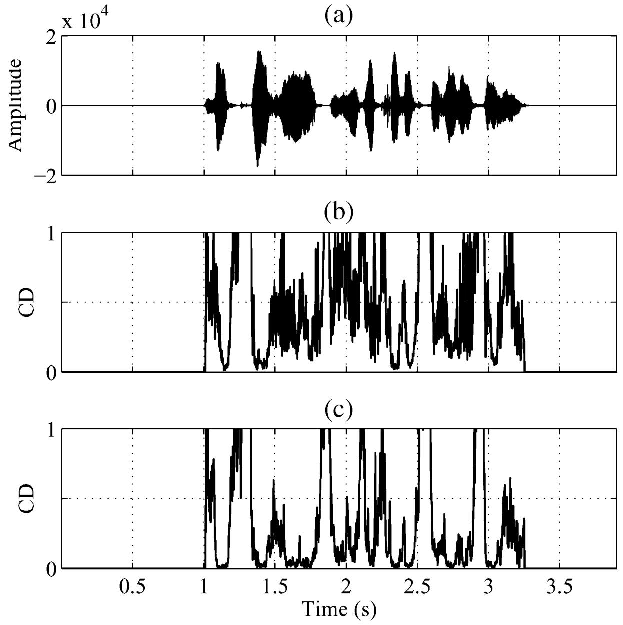

Fig. 11. (a) Clean speech and cepstral distances (CD) between clean speech and (b) speech enhanced by TSNR technique and (c) speech enhanced by HRNR technique.

Table I generalizes the previous example for a speech database lasting 72 min. This corpus is composed of four speakers (two females and two males), nine sentences per speaker, five SNR conditions (0, 6, 12, 18, and 24 dB) and three noise types (Street, Car, and Babble). The input SNRs are computed using the ITU-T recommendation P.56 [14] speech voltmeter (SV56). Table I presents values obtained for TSNR and HRNR techniques, the CD being computed between clean speech and enhanced speech. For each sentence, the CD values are averaged during speech activity giving a mean CD. For each noise type and SNR value, a mean CD is given that is the result of the averaging of the mean CD obtained for 36 sentences. The proposed HRNR technique achieves the best results (bold values) under all noise conditions which confirms that this approach succeeds in limiting speech degradations introduced by TSNR. These degradations are mainly due to the noise PSD estimation errors inherent to single channel speech enhancement techniques. However, the HRNR technique is able to overcome this limitation for voiced speech components by regenerating the degraded harmonics in order to compute a spectral gain preserving these harmonics. However, when the input SNR is too small, i.e., 0 dB, the improvement is small which confirms the analysis of

TABLE I
MEAN CEPSTRAL DISTANCE BETWEEN CLEAN SPEECH AND SPEECH ENHANCED USING TSNR AND HRNR TECHNIQUES, RESPECTIVELY, FOR VARIOUS NOISE TYPES AND SNR CONDITIONS

Noise type	Input SNR (dB)	Mean Cepstral Distance	
		TSNR	HRNR
Street	0	1.05	**1.00**
	6	0.90	**0.81**
	12	0.75	**0.61**
	18	0.58	**0.44**
	24	0.42	**0.31**
Car	0	0.89	**0.85**
	6	0.75	**0.62**
	12	0.60	**0.44**
	18	0.46	**0.32**
	24	0.37	**0.22**
Babble	0	1.09	**1.03**
	6	0.89	**0.79**
	12	0.71	**0.58**
	18	0.52	**0.40**
	24	0.35	**0.25**

TABLE II
OUTPUT AVERAGE SEGMENTAL SNRS USING TSNR AND HRNR TECHNIQUES IN VARIOUS NOISE AND SNR CONDITIONS

Noise type	Input SNR (dB)	Average segmental SNR (dB)	
		TSNR	HRNR
Street	0	3.44	**3.67**
	6	8.16	**8.79**
	12	13.31	**14.16**
	18	18.58	**19.41**
	24	23.95	**24.62**
Car	0	4.97	**5.31**
	6	9.28	**9.93**
	12	14.01	**14.84**
	18	18.91	**19.72**
	24	24.04	**24.74**
Babble	0	3.42	**3.69**
	6	7.91	**8.53**
	12	13.29	**14.20**
	18	19.03	**20.02**
	24	24.78	**25.71**

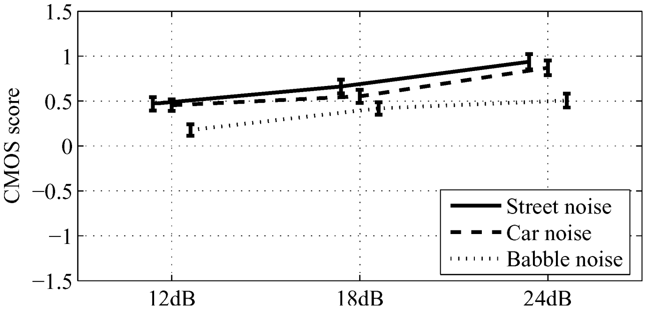

Fig. 12. Results of the CCR test between TSNR and HRNR algorithms. CMOS scores and confidence intervals are given for three SNRs (12, 18, and 24 dB) and three noises types (street, car, and babble).

Section VI-B. Actually, in such a condition, the TSNR approach cannot restore enough harmonics to make the harmonic regeneration process efficient.

Based on the database described in the previous paragraph, Table II presents the input SNRs of noisy speech and the corresponding average segmental SNRs obtained using TSNR and HRNR techniques. The segmental SNR measure takes into account both residual noise level and speech degradation and can be computed, during speech activity, as follows:

$$\text{segSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\sum_{l=Lm}^{Lm+L-1} s^2(l)}{\sum_{l=Lm}^{Lm+L-1}(\hat{s}(l)-s(l))^2} \qquad (39)$$

where M is the number of frames that contain active speech, and l is a discrete-time index. For each noise type and SNR value, the average segmental SNR is the result of the averaging of the segmental SNRs obtained for 36 sentences. The HRNR technique achieves the best results (bold values) under all noise conditions. The segmental SNR improvement brought by the HRNR technique is explained by its ability to preserve the harmonics degraded by the TSNR.

B. Formal Subjective Test

To confirm the objective results, a formal subjective test has been conducted. It consists in a comparative category rating (CCR) test compliant into the UIT-T recommendation P.800 [15]. For each algorithm, TSNR and HRNR, the parameters have been tuned to obtain a satisfactory tradeoff between noise reduction and speech distortion. The 0- and 6-dB SNR levels were judged too critical and then were not retained in this subjective test. This test was conducted with 24 listeners and using the corpus described in Section VII-A. The listeners had to listen the sentences by pairs (TSNR technique—HRNR technique or in reverse order, the order being random) and then rate the second sentence in contrast to the first one. The scale goes from -5 to 5 by steps of 1. The listeners used this scale to give global preference that take into account both residual noise level and distortion level. The results obtained are displayed in Fig. 12. The comparative mean opinion score (CMOS) and the associated confidence interval are displayed versus the SNR for each noise type. A positive value indicates that the HRNR technique is preferred to the TSNR one. We can observe that the HRNR technique is always preferred, with significant mean scores, to the TSNR technique which is in agreement with the objective results presented in Tables I and II. However, there is less improvement for the babble noise (speech-like noise) than for street and car noises. This is recurrent for speech enhancement techniques as it is difficult to deal with nonstationary noises. We can also note that the amelioration increases with the SNR. As explained in Section VI-B, the efficiency of the HRNR technique depends on the degradation level of the signal. It is easier to restore harmonics when only a few are degraded or missing which explains the better behavior for high SNRs.

VIII. CONCLUSION

In this paper, we have proposed and analyzed a noise reduction technique in order to improve the DD approach. The TSNR technique is based on the estimation of the *a priori* SNR in two steps. The *a priori* SNR estimated using the DD approach shows interesting properties but suffers from a frame delay which is removed by the second step of the TSNR algorithm. So, this technique has the ability to immediately track the nonstationarity of the speech signal without introducing musical noise. Consequently, the speech onsets and offsets are preserved and the reverberation effect characteristic of the DD approach is removed.

We have also proposed a noise reduction technique based on the principle of harmonic regeneration. Classic techniques, including the TSNR, suffer from harmonic distortions when the SNR is low. This is mainly due to estimation errors introduced by the noise PSD estimator. To solve this problem, a nonlinearity is used to regenerate the degraded harmonics of the distorted signal in an efficient way. The resulting artificial signal helps to refine the *a priori* SNR which is then used to compute a spectral gain that preserves speech harmonics, and hence avoids distortions. The role of the nonlinearity and the principle of harmonic regeneration have been detailed and analyzed. Results are given in terms of cepstral distance and segmental SNR on a large corpus of signals to illustrate the efficiency of the HRNR technique. All these results demonstrate the good performance of the HRNR technique in terms of objective results. For the sake of completeness, results of a formal subjective test have been given and confirm the significant performance improvement brought by the HRNR technique.

References

[1] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, May 1996, vol. 2, pp. 629–632.

[2] Y. Ephraïm and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraïm and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[4] C. Plapous, C. Marro, P. Scalart, and L. Mauuary, "A two-step noise reduction technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montréal, QC, Canada, May 2004, vol. 1, pp. 289–292.

[5] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 157–160.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwith compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[8] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[10] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1984, vol. 9, pp. 53–56.

[11] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[12] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. Workshop Consistent Reliable Acoust. Cues Sound Anal.*, Aalborg, Denmark, Sep. 2001, pp. 71–74.

[13] R. F. Kubichek, "Standards and technology issues in objective voice quality assessment," *Digital Signal Process.*, vol. 1, pp. 38–44, 1991.

[14] Telephone Transmission Quality—Objective Measuring Apparatus Geneva, Switzerland, Mar. 1996, ITU-T Rec. P.56.

[15] Methods for Subjective Determination of Transmission Quality Geneva, Switzerland, Aug 1996, ITU-T Rec. P.800.

Cyril Plapous (M'04) was born in Lannion, France, in 1979. He received the "Diplôme d'Ingénieur" degree from École Nationale Supérieure de Sciences Appliquées et de Technologie (ENSSAT), Lannion, France, and the "Diplôme d'Études Approfondies" (M.S.) degree in signal, telecommunication, image, and radar from the University of Rennes, in 2002. He is currently working toward the Ph.D. degree at France Télécom R&D/TECH/SSTP, Lannion, France, in the field of speech enhancement.

He was Trainee at ATR Adaptive Communications Research Laboratories, Kyoto, Japan, in 2002.

Claude Marro was born in Nice, France, in 1967. He received the Ph.D. degree in signal processing and telecommunications from the University of Rennes, Lannion, France, in 1996.

He worked on speech dereverberation and noise reduction using multimicrophone techniques for interactive communication applications. Since 1997, he has been with France Télécom R&D, Lannion, as a Research Engineer in acoustics and speech signal processing. His current research interests include speech enhancement, echo cancellation, and voice modification applied to communication and multimedia contexts.

Pascal Scalart (M'06) received the Ph.D. degree in signal processing and telecommunications from the University of Rennes, Lannion, France, in 1992.

In 1993, he held a Postdoctoral position at Laval University, Québec, QC, Canada, engaging in research on digital signal processing for communications. Since 1994, he has been with France Télécom R&D, Lannion, where he has been involved in research on speech signal processing for multimedia applications in the field of speech enhancement and adaptive filtering techniques for echo cancellation. He is currently a Professor at the University of Rennes and is a member of the research laboratory IRISA.